ESTIMATION DE LA CHARGE À L'ULTIME POUR LES ÉVÈNEMENTS CLIMATIQUES DE GRANDE AMPLEUR

Mouhamed Moustapha NDOUR

07/11/2023

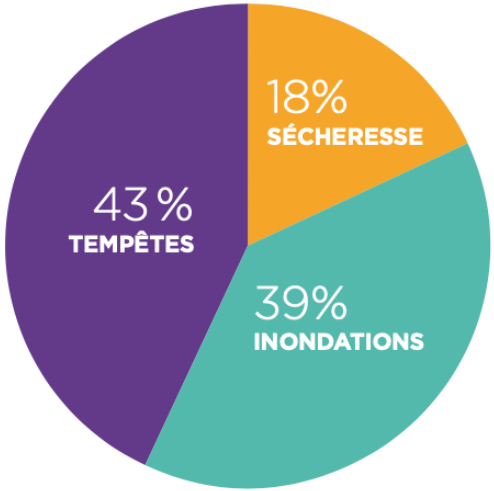
4, rue Chauveau-Lagarde – 75008 Paris – tél. 01 44 51 72 72
www.institutdesactuaires.com

• SOMMAIRE

• SOMMAIRE

- 1- Contexte et objectif
  Quelques chiffres clés – France

Répartition du cumul des indemnisations versées par les assureurs au cours des 31 dernières années (1989 – 2019)



Différents aléas climatiques

- Inondation : **28,8 Mds €** d'indemnisations cumulées entre 1989 et 2019
- Tempête : **31,6 Mds €** d'indemnisations cumulées entre 1989 et 2019
- Sécheresse : **13,8 Mds €** d'indemnisations cumulées entre 1989 et 2019

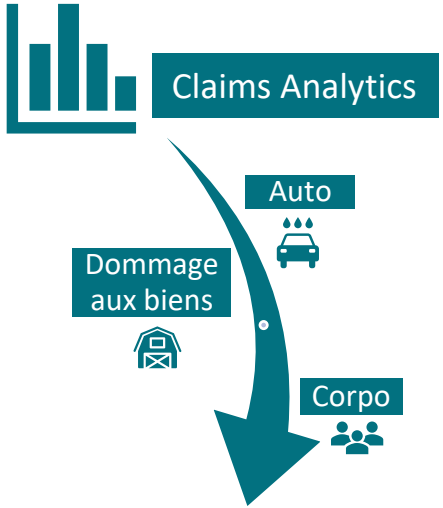
Impacts climatiques

De 1989 à 2019	Nombre de sinistres indemnisés	Charge (Md€ constants 2020)
INONDATIONS	**1 961 000**	**28,8**
Particuliers	1 480 000	15,1
Professionnels	481 000	13,6
TEMPÊTES	**10 105 000**	**31,6**
Particuliers	8 251 000	17,9
Professionnels	1 854 000	13,7
SÉCHERESSE	**843 000**	**13,8**
Ensemble des périls	**12 909 000**	**74,1**

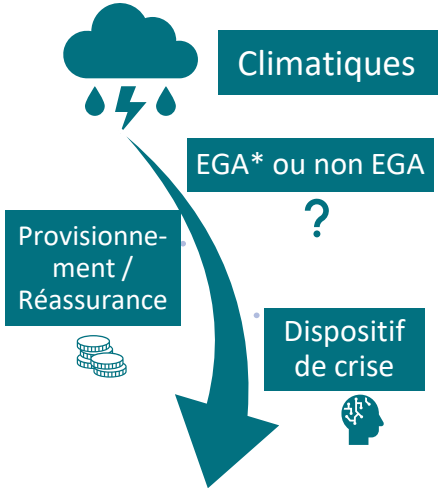
Source : Etude Climat France Assureurs Octobre 2021

- 1- Contexte et objectif
  Périmètre d'étude et enjeux

Toutes les activités d'assurance IARD



Maîtrise du coût moyen



Prédiction de la charge à l'ultime pour les EGA

* EGA : Evènements de Grande Ampleur pour lesquels la charge brute à l'ultime >1M€

1- Contexte et objectif

Objectif du mémoire

→ Gagner en précision, fiabiliser et automatiser la méthode d'estimation de la charge à l'ultime pour les EGA

⚠ Changement du seuil de détection des EGA (passage de 3M€ à 1M€)

Approche AXA France

- − *Base de données incomplète suite au changement de seuil de détection des EGA*
- − *Méthode manuelle avec une estimation basée sur la visualisation des évènements les plus proches → Difficulté à trouver des évènements comparables*
- − *Pas de prise en compte de l'aspect géographique*

Approche proposée

- + *Base de données plus exhaustive, gain en robustesse statistique*
- + *Méthode automatisée prenant en compte plus de variables explicatives (notamment géographiques)*
- + *Gain en fiabilité*

Construction d'une nouvelle base de données EGA

Estimation de la charge par Machine Learning et GLM

• SOMMAIRE

• 2- Définition d'un EGA et méthode d'estimation de la charge
Définition d'un EGA

Méthode actuelle

- Les **EGA** représentent les **E**vènements de **G**rande **A**mpleur, c'est à dire les évènements climatiques qui ont une charge élevée (supérieure à un seuil fixé)

- Actuellement, on définit les EGA en ne considérant que les évènements dont la **charge est supérieure à 1M** (3M avant 2021)

- Mise en place d'un partenariat entre AXA et Predict (filiale de Meteo France) qui envoie des alertes avec un indice de gravité par commune lorsqu'un évènement pourrait avoir lieu mais également pour prévenir les assurés

Méthode d'estimation de la charge (Modèle AXA)

→ Alerte Predict

→ Construction des premières cadences d'ouverture des sinistres

Retrait des week-ends et jours fériés

→ Choix de l'évènement le plus proche de celui à estimer

Calcul de la somme de la différence des carrés des matrices de **cadencement du <u>nombre</u> d'ouvertures**

$$\text{Distance} = \sum_{i,j=1}^{i=n,\ j=m} \left({x_{ij}}^2 - {y_{ij}}^2\right)$$

→ Estimation du volume de sinistres à l'ultime

$$\text{Volume ultime} = \frac{Volume\ sinistres\ \text{à}\ JO+i}{\%\ du\ volume\ ultime\ de\ l'évènement\ le\ plus\ proche\ \text{à}\ JO+i}$$

→ Estimation du coût moyen hors graves à partir des évènements passés plus récents et de même nature + ajout du coût moyen des sinistres graves transmis par l'inspection

→ Estimation de la charge à l'ultime

Charge à l'ultime = Volume ultime ∗ Coût moyen

Méthode d'estimation de la charge (Ex:Tempête Eunice)

- Survenance : du 17/02 au 19/02/22
- Régions les plus impactées : Pas-de-Calais et zones limitrophes, Manche, Seine-Maritime

Alerte Predict

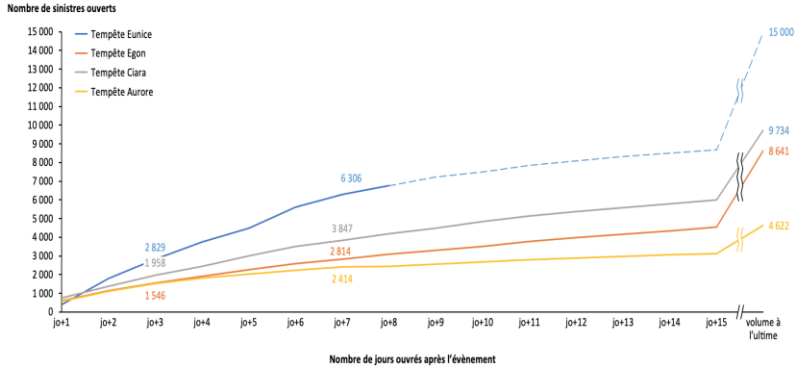
Chiffres sinistralité

vision au 26 février 2022	
Nombre de sinistres	charge d/d
6 306	18,5 M€

projection à l'ultime	
Nombre de sinistres	Charge à l'ultime
12 500	34,7 M€

On compare les cadences d'ouverture des sinistres précédents de même type et on garde celles qui se rapprochent le plus de la tempête Eunice. Il s'agit ici des tempêtes Egon, Ciara et Aurore.
La tempête Ciara a le cadencement du nombre d'ouvertures le plus proche de celle que l'on cherche à estimer.
Une première estimation le 23 février entre 13 et 18k sinistres.
A JO+30, 11500-13000 sinistres pour une charge à l'ultime de 31 à 35 M€ à date.

Evolution des volumes de sinistres sur des tempêtes comparables à Eunice



• SOMMAIRE

• 3 - Construction de la nouvelle base de données EGA

Objectif : Consolidation de l'identification des Evénements de Grande Ampleur au sein des données climatiques

→ Créer un nouveau catalogue, une nouvelle base d'apprentissage consolidée et fiable afin de l'exploiter et de prédire la charge ultime

2 raisons principales :

→ Passage du seuil de détection des EGA de 3M€ dans le passé à 1M€ aujourd'hui

→ Possibilité d'omission de plusieurs évènements ou intempéries plus ou moins diffus avant le partenariat avec Predict

Grandes étapes :

Lecture et nettoyage des données

Lecture

- → Base de données contenant l'ensemble des UP* climatiques ouvertes entre 1989 et 2021 (près de 3 millions de lignes)
- → Prise en compte des dimensions géographique et temporelle ainsi que de la variable cible « Y » qui est le volume d'UP, c'est-à-dire le nombre de garanties ouvertes suite aux sinistres

Nettoyage

- → Près de 1,5 millions de lignes ne contiennent pas d'info sur le lieu du sinistre
- → Certaines lignes peuvent être existantes mais l'information renseignée est inexploitable.

Ex: DTM_Lieu_Du_Sinistre == 'RES.'

- → On commence par dissocier les codes postaux des lieux de sinistres contenus dans la variable puis on utilise les données externes afin de mapper les codes postaux avec un nom de commune et vice-versa

Ex: si on dispose du code commune 59350, on va récupérer comme lieu du sinistre LILLE.

UP : Unité de Prestation, garantie liée au sinistre

Lecture et nettoyage des données

→ Exploitation des données contrat pour récupérer le plus d'informations relatives au lieu possible. On se retrouve au final avec près de 70% des lignes avec une information géographique clairement renseignée

→ Parmi les lignes non renseignées, plus de 80% sont des événements de 1989 à 1999.

→ On va filtrer les données sur l'horizon 2000-2021 en ne gardant que les données de ce scope qui contiennent une information géographique exploitable.

→ On se retrouve désormais avec un peu plus 2 240 000 lignes allant de 2000 à 2021

Examen préliminaire

⚠ On observe une trop grande disparité entre les communes au cours de ces années d'historique en termes de volume ainsi qu'en termes de fréquence journalière

→ Création d'un zonier afin "d'harmoniser" les volumes d'UP associé à chaque individu (on passe donc d'une union de communes à une union de zones paramétriques

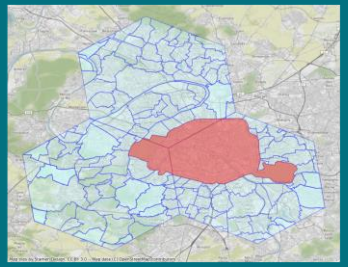
Institut des ACTUAIRES

Découpage spatial

Découpage de la France de manière géométrique et équitable.

Exemple : Paris

Paris apparait en rouge. La ville est à cheval et fait partie de trois zones hexagonales (en bleu).

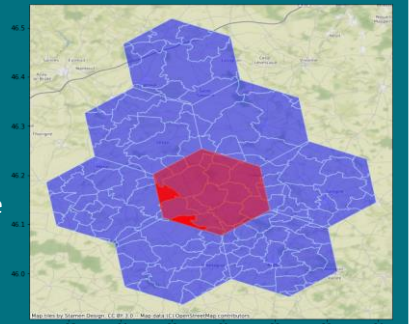


Identification des zones à traiter et qui ne se suffisent pas à elle-même

On utilise pour cela une variable de seuil (= volume global d'UP observé à LILLE entre 2000 et fin 2021).

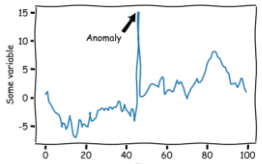
Exemple : Commune de REMIREMONT

La commune de REMIREMONT est localisée au sein de la zone rouge; la zone agrégée créée autour d'elle et pour laquelle elle appartient apparaît en violet.
7 voisins ont été nécessaires afin d'atteindre la valeur de seuil S



DTSURV	hex_id_unique_commune	total_volume_up	total_merlin_chg	lieu_du_sinistre	code_postal_actualise	code_commune_actualise	commune_ou_commune_de_rattachement	zone_etendue	CATNA_volume_up	CATNA_merlin_chg
2018-02-02	851fb00bfffffff	14.500000	60.641040	VILLENEUVE LE ROI, SAINT BRICE, ESBLY, MELUN, ...	77400, 94210, 94290, 77920, 77450, 77000, 7716...	77288, 77243, 94077, 94068, 77125, 77441, 7740...	VILLENEUVE LE ROI, SAINT BRICE, ESBLY, SAINT M...	851fb053fffffff, 851fb06fffffff, 851fb087fff...	8.333333	45.17081
2018-01-03	851fb13bfffffff	31.833333	91.056282	BETTANCOURT LA FERREE, THORS, VILLIERS EN LIEU...	10200, 51290, 10330, 51340, 52130, 52330, 5211...	10378, 51583, 51080, 52411, 10192, 52182, 5211...	BETTANCOURT LA FERREE, THORS, VILLIERS EN LIEU...	851fb107fffffff, 851fb12bfffffff, 851fb133fff...	0.000000	0.00000
2018-04-29	851fa647fffffff	18.500000	33.094193	CHARLEVILLE MEZIERES, REVIN, GIVRY, YVERNAUMON...	08310, 08090, 08130, 08500, 51100, 02140, 0800...	08076, 08193, 08361, 08338, 02341, 08105, 0800...	CHARLEVILLE MEZIERES, REVIN, GIVRY, YVERNAUMON...	851fa653fffffff, 851fb583fffffff, 851fa2dbfff...	0.000000	0.00000

Modélisation

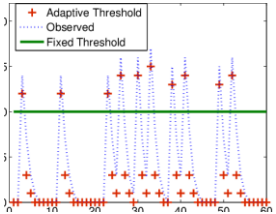


→1ere approche: Threshold Models

- Une première modélisation se base sur le dépassement de seuil
- Un EGA pourra donc être vu comme **une somme d'anomalies** *("positive", un maximum local)* au sein d'une série temporelle, c'est-à-dire un événement inhabituel qui semble s'éloigner, dépassant un certain seuil et ne pas se comporter comme les autres événements

Fixed Threshold Model : Modèle basé sur une variable de seuil fixe

Adaptive Threshold Model : Modèle paramétrable basé sur une variable de seuil adaptative



Choix du seuil : Théorie des valeurs extrêmes

→ Loi GPD et théorie des valeurs extrêmes

$$G_{\sigma\xi}(y) = \begin{cases} 1 - (1 - \xi\frac{y}{\hat{\sigma}})^{-\frac{1}{\xi}}, \ \forall\ \xi \neq 0 \\ 1 - exp(-\frac{y}{\hat{\sigma}}), \ si\ \xi = 0 \end{cases}$$

- Si $\xi < 0,\ y \in [0, Min(-\frac{\sigma}{\xi}, x_F - u)]$
- Si $\xi \geq 0,\ y \in [0, , x_F - u]$

où $G_{\sigma,\xi}$ la fonction de répartition de la loi de Pareto généralisée (GPD), $\xi \in R$ le paramètre de forme et $\hat{\sigma} > 0$ le paramètre d'échelle.

→ Choix du seuil

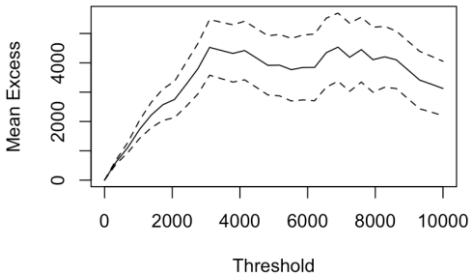
-Durée de vie résiduelle moyenne

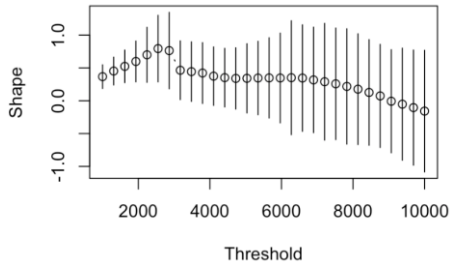
Soit $X - u_0 | X > u_0 \sim GPD(\sigma, \xi)$, $\xi < 1$, alors $\forall u \geq u_0$

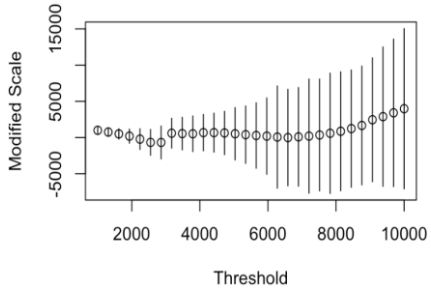
$$MRL(u) = \mathbb{E}(X - u | X > u) = \frac{\sigma_{u_0} + \xi_u}{1 - \xi}$$

-Stabilité par seuil du paramètre d'échelle

-Stabilité par seuil du paramètre de forme

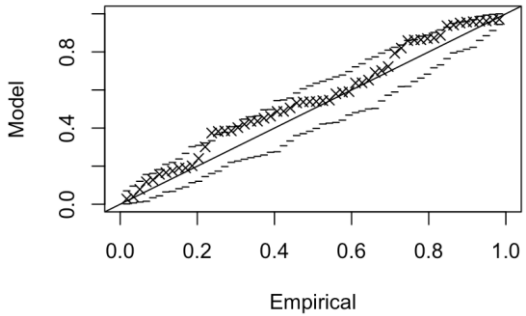




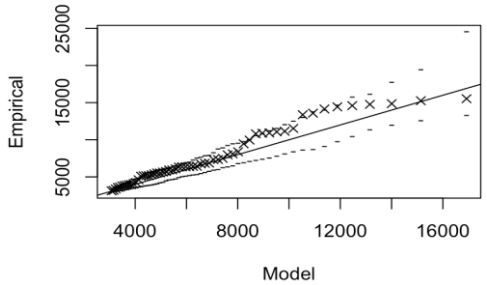


→ Seuil entre 3000 et 4000

→ Seuil autour de 3000



→ Probability plot



→ Quantile plot

2ème approche: CART (Classification And Regression Trees) et l'algorithme d'Isolation Forest

En entrée : Soient les variables aléatoires X1, ..., Xn indépendantes, ε_Φ un échantillon aléatoire de X de taille Φ

En sortie : On obtient un groupe d'arbres binaires.

Si ε_Φ ne peut pas être divisé :

Alors retourner un noeud externe de taille Φ

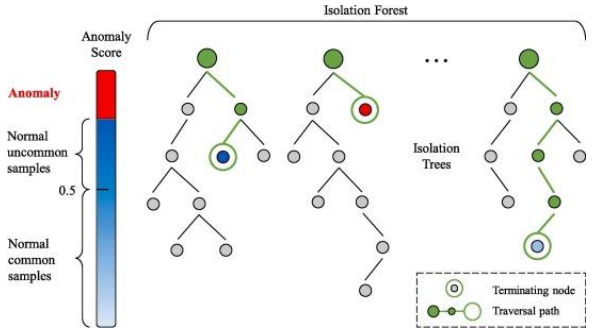
Sinon :

Sélectionner une variable aléatoirement

Découper cette variable aléatoirement selon un seuil de séparation (toute valeur dans la plage des valeurs minimum et maximum de la variable sélectionnée).

Répéter les deux étapes précédentes jusqu'à l'isolation d'une donnée.

Répéter les étapes précédentes de façon récursive

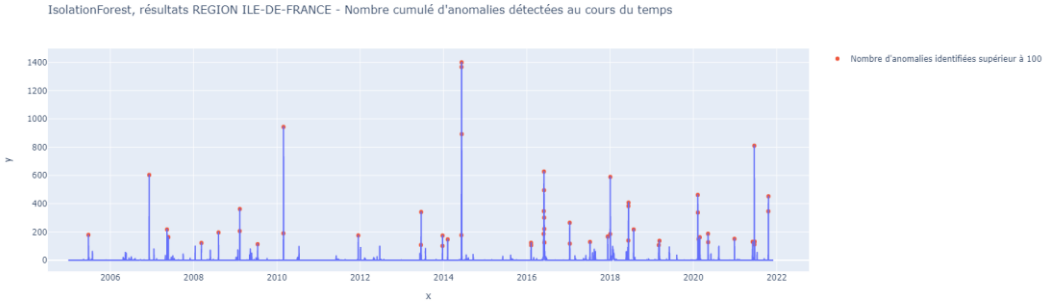


Modélisation

Résultats

Modèle	Nombre total d'événements identifiés
Fixed T.	68%
Adaptive T. (1%)	81%
Adaptive T. (3%)	78%
Adaptive T. (5%)	75%
IsolationForest	93%

Résultat Isolation Forest : Nombre cumulé d'EGA détectés au cours du temps en Ile de France



Aperçu du nouveau catalogue à date (avec Isolation Forest)

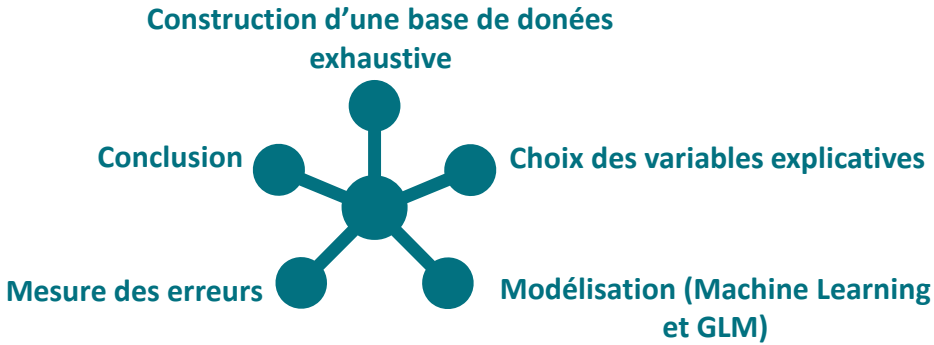
DTSURV	EGA_isolationForest	code_commune_actualise	lieu_du_sinistre	code_postal_actualise	commune_ou_commune_de_rattachement	occurrence_up	merlin_chg
2017-07-10	NaN	71550	UCHIZY	71550	UCHIZY	8	27.04973
2020-10-20	NaN	43140	BLAVOZY	43140	LE MONTEIL	1	0.50010
2006-12-08	1.0	45137	ESCRENNES	45137	ESCRENNES	2	0.00000
2000-11-28	NaN	42095	FIRMINY	42095	FIRMINY	1	0.00000
2014-11-27	NaN	86066	CHATELLERAULT	86066	CHATELLERAULT	1	2.15700

• SOMMAIRE

- 4- Modèles de projection

Objectif : Appliquer des méthodes statistiques plus robustes que les méthodes graphiques actuellement utilisées à AXA

Différentes étapes:



Choix des variables explicatives (principales)

→DTOUV : Date d'ouverture des sinistres

→ zones_agregees_dappartenance

→ part_auto : Proportion de sinistres liés à l'auto

→ part_dommages : Proportion de sinistres liés aux dommages

→ part_transport : Proportion de sinistres liés au transport

→ DTSURV : Date de survenance des sinistres

→ part_constructions : Proportion de sinistres liés aux constructions

→ Part_RC : Proportion de sinistres liés à la RC

→ dureeEvenement : Durée de l'évènement climatique

→ "J1", "J2" et "J3" : 3 premières cadences d'ouverture

Modélisation
Estimation du volume de sinistres par Machine Learning

1ère étape : Construction des cadences d'ouvertures des sinistres

→ On détermine les cadences d'ouverture en prenant les sommes cumulées sur les nombres de sinistres comptabilisés.

→ Lorsque des évènements surviennent le week-end, les ouvertures sont comptabilisées le lundi ou le lendemain lorsqu'il s'agit de jours fériés.

Exemple:

evt_new	debut	fin	Type_even	DTSURV	dureeEvenement	DTOUV	JourOuv	nbresin
1.0	2009-01-23	2009-01-25	Tempete	2009-01-23	2 days	2009-01-23	Friday	53
						2009-01-24	Saturday	9
						2009-01-26	Monday	139
						2009-01-27	Tuesday	131
						2009-01-28	Wednesday	101

→

JourOuv	cumsum	Indic	Cadence
Friday	53.0	1	53.0
Monday	201.0	1	201.0
Tuesday	332.0	1	332.0
Wednesday	433.0	1	433.0
Thursday	500.0	1	500.0

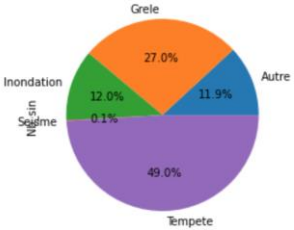
→

J1	J2	J3	JFinal
201.0	332.0	433.0	1605.0
3119.0	6931.0	10521.0	79286.0
115.0	162.0	180.0	494.0
128.0	253.0	383.0	3192.0
1157.0	1657.0	2154.0	10905.0

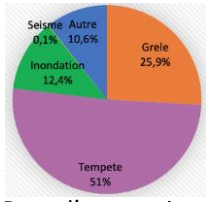
Modélisation

Estimation du volume de sinistres par Machine Learning

2ème étape : Décomposition de la base de données en base d'apprentissage et base de test (70%-30%)



Base initiale



Base d'apprentissage

→Algorithmes :

- KNN : K plus proches voisins
- Forêts aléatoires
- Gradient Boosting

→Mesures d'erreur

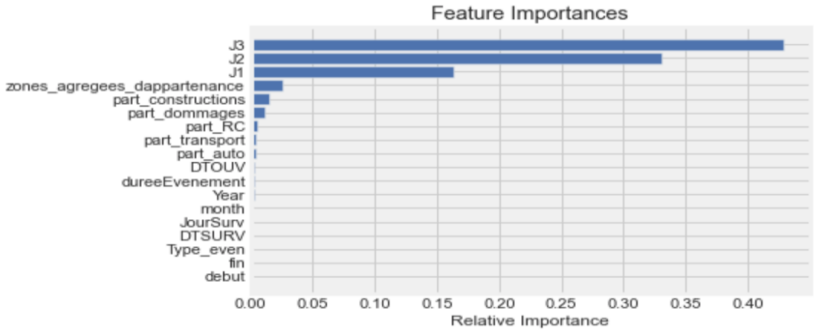
$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

	KNN	Random Forest	Gradient Boosting
RMSE	6287.7403	5827.0283	5681.2081
MAPE	0.6150	0.2503	0.2253

→ Meilleur modèle : gradient boosting qui a le RMSE le plus faible et le MAPE le plus faible (~22%)

→ L'importance des variables permet de déduire la contribution de la variable géographique dans la construction du modèle



→Les variables des cadences sont celles à plus forte contribution au niveau du modèle Gradient Boosting.

→La zone géographique arrive en 4ème position

Modélisation

Estimation du coût moyen par Modèle Linéaire Généralisé (GLM)

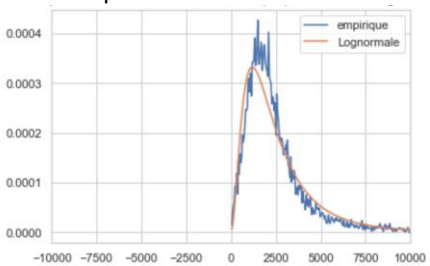
- Lois utilisées :
  - →Log-normal
  - →Gamma

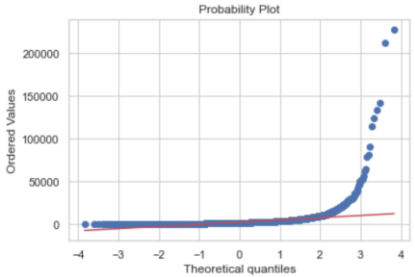
- Critères pour le choix de la loi
  - →Déviance : caractérisée par une variation de la log-vraisemblance
  - →Critère AIC = $-2\log(L) + 2p$
  - →Critère BIC = $-2\log(L) + p.\log(n)$

  où *log(L) est la log-vraisemblance, p le nombre de paramètres et n le nombre d'individus dans l'échantillon*

Résultats modèles linéaires généralisés

→Tempête:





	Déviance	Critère AIC	Critère BIC
Lognormale	56445.6696	57994.1779	58000.9267
Gamma	57451.0899	58873.3999	58880.1487

→Modèle retenu : Lognormal

→ Grêle, inondation et séisme : mêmes résultats

→ Evaluation de la performance du modèle choisi

	Tempête	Grêle	Inondation	Séisme
MAPE	0.0957	0.0937	0.0965	0.0971

Coût moyen prédit avec un intervalle de confiance de +/- 10% pour les 4 risques.

• SOMMAIRE

- 5- Comparaison des modèles

	Modèle AXA	Modèle ML/GLM
Base de données	Moins exhaustive	Plus exhaustive
Variables explicatives utilisées	Typologie d'évènement Dates de survenance Cadences d'ouverture	+ Zones géographiques, Types de client , Branches (Auto, DAB…)
Marge d'erreur moyenne pour l'estimation du volume des sinistres	15%	22%
Marge d'erreur moyenne pour l'estimation du coût moyen	15%	10%
Projection volume de sinistres Eunice	15700 (13 500 – 18 000) à JO+3 11 500 – 13000 à JO+30	10800 (8000 – 13 000) à JO +3
Projection de la charge finale prévisible Eunice	42 M€ (35 – 50 M€) à JO +3 31- 35 M€ à JO +30	31,5 M€ (23 M€ – 39 M€) à JO+3

Charge à date pour Eunice: 31 M€
Volume à date pour Eunice: 11 000

• SOMMAIRE

- Conclusion

RESULTATS

→ Prise en compte de plus de variables pertinentes

→ Estimation plus précise du coût moyen

→ Estimation plus fiable du volume

→ Estimation globalement plus fiable de la charge

LIMITES

→ Restriction de notre base à l'horizon 2000-2021 en raison des contraintes liées au manque d'informations géographiques

→ Pas de prise en compte de l'évolution du portefeuille dans notre étude

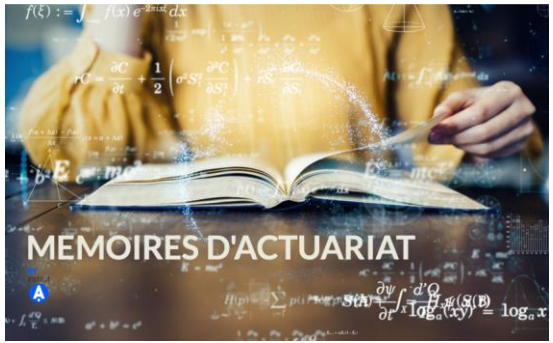
→ Utilisation des mêmes variables explicatives pour les différents aléas climatiques

AXES D'EVOLUTION

→ Test des modèles sur de nouveaux évènements climatiques dont la sévérité est bien plus importante

→ Test des modèles sur d'autres aléas climatiques (Ex: Sécheresse)

→ Prise en compte de l'évolution du portefeuille



MERCI POUR VOTRE ATTENTION!